Two-Stage Designs versus European Scaled Average Designs in Bioequivalence Studies for Highly Variable Drugs: Which to Choose?

Eduard Molins[a*†], Erik Cobo[a], Jordi Ocaña[b]

Abstract

The usual approach to determine bioequivalence for highly variable drugs is scaled average bioequivalence, which is based on expanding the limits as a function of the within-subject variability in the reference formulation. This requires separately estimating this variability, and thus using replicated or semi-replicated crossover designs. On the other hand, regulations also allow using common 2×2 crossover designs based on two-stage adaptive approaches with sample size re-estimation at an interim analysis. The choice between scaled or two-stage designs is crucial and must be fully described in the protocol. Using Monte Carlo simulations, we show that both methodologies achieve comparable statistical power, though the scaled method usually requires less sample size, but at the expense of each subject being exposed more times to the treatments. With an adequate initial sample size (not too low, e.g., 24 subjects), two-stage methods are a flexible and efficient option to consider: They have enough power (e.g., 80%) at the first stage for non-highly variable drugs and, if otherwise, they provide the opportunity to step up to a second stage that includes additional subjects.

[a]Department of Statistics and Operations Research, Universitat Politècnica de Catalunya.
[b]Department of Genetics, Microbiology and Statistics. Universitat de Barcelona.
[*]Correspondence to: Eduard Molins, Department of Statistics and Operations Research, Universitat Politècnica de Catalunya. Jordi Girona 1-3, 08034 Barcelona, Spain.
[†]E-mail: eduard.molins@astrazeneca.com

Include up to six keywords that describe your paper for indexing purposes: Average Bioequivalence (ABE), Reference Scaled Average Bioequivalence (RSABE), Two-Stage Designs (TSD), Highly Variable Drugs (HVD), Significance Level Adjustment

---

1. Introduction

Average bioequivalence (ABE) studies are conducted to demonstrate in vivo either that two products, say "test" *T* and "reference" *R*, are pharmaceutically equivalent (in the US) or that their rate and extent of absorption[1-3] are close enough to serve as alternative pharmaceutical products (in the EU). The most common measure of the rate of absorption is the bioavailability measure "maximum observed concentration" (*Cmax*), while the "area under the concentration curves" ($AUC_{0\text{-}t}$ and $AUC_{0\text{-}\infty}$)[4] are the most common bioavailability measures for the extent of absorption. To demonstrate ABE, regulatory guidelines recommend a single dose 2×2 crossover design, *RT/TR* that evaluates *T* and *R* on healthy volunteers. The most commonly used criterion to test (at a significance level of $\alpha = 0.05$) for ABE is the "interval inclusion rule", which is based on a 90% symmetric confidence interval for the formulation effect, say the mean difference between the bioavailabilities of formulations *T* and *R* at a log-transformed

scale. It is based on the Student's distribution, assuming data normality. In order to declare ABE, the back-transformed confidence interval for the geometric means ratio (*GMR*) should lie fully within the ABE limits of 0.80–1.25 (=1/0.80), corresponding to ±0.223 on the logarithmic scale.[2,5]

Highly variable drugs (HVD) are characterized by high within-subject variability in the rate and/or extent of absorption of its active principle. This hinders researchers from declaring ABE when it really holds, unless unacceptably large sample sizes are used. Most regulations classify a drug as HVD if the within-subject coefficient of variation of the reference formulation *R* (CV_{WR}) is 30% or greater on the original scale. The percentage of HVD is not negligible. Davit et al[6] collected data from all in vivo bioequivalence studies reviewed by the FDA's Office of Generic Drugs from 2003 to 2005, and they concluded that 31% of the studies (57/180) corresponded to highly variable drugs, many of them around CV_{WR} = 30%.

If HVD is suspected, the European Medicines Agency (EMA) allows linearly scaling the *Cmax* margins as a function of the *R* variability to a maximum plateau of 0.6984-1.4319, and it further allows application of the interval inclusion rule over these expanded limits.[2] Similarly, the FDA also allows researchers to re-scale the *AUC* limits.[1,3] These scaled approaches require the use of high order crossover designs like the replicated *TRTR/RTRT* or semi-replicated *TRR/RTR/RRT* designs.[2,7,8] However, these scaled methods, as defined by FDA and EMA regulations, do not adequately preserve the type I error rate in the neighborhood of CV_{WR} = 30%.[9,10] Thus, the proportion of non-ABE products erroneously declared as ABE is higher than its desired nominal value.

Regulators also allow using two-stage adaptive designs (TSD) with unblinded interim sample size re-estimation[2,5,11,12] based on the usual 2×2 crossover *RT/TR* design. Bioequivalence may be declared at the interim look with N_1 subjects; otherwise, the sample size can be increased on the basis of the estimated within-subject variability at the first stage, then ABE is tested again at a second stage with cumulated data $N = N_1 + N_2$. Two-stage designs preserve the type I error rate[13] by adjusting significance boundaries at each stage in various ways that are not fully specified in the regulations.[14,15]

In turn, the planned sample size is crucial because it may lead to underpowered studies, as there is a high uncertainty about the assumed *GMR* and/or variability.

The main objective of this paper is to critically compare the EMA's original scaled method based on a replicate *TRTR/RTRT* design (or, more precisely, an adjusted variant intended to preserve the type I error rate, as shown by Labes and Schütz[10]) with two TSD methods based on the usual *RT/TR* crossover design.

Section 2 describes the compared methods and details the simulation methodology. Section 3 shows the results; and Section 4 discusses them in order to recommend the most appropriate approach.

2. Statistical methodology

2.1. 2010 Regulatory EMA reference scaled average bioequivalence approach (RSABE) (for Cmax only)

Replicate *TRTR/RTRT* designs allow separately estimating the CV_{WR} [9] and can easily be re-arranged for comparison with a 2×2 crossover design (needed for two-stage designs) once the first two periods are sliced (see Section 2.3).

We focus on the EMA regulation because the FDA's approach is based on scaled limits which are discontinuous at CV_{WR} = 30%. This discontinuity is associated with a sharp peak of type I probability around this *CV* value, which threatens its validity.

On the original scale, the null hypothesis of bioinequivalence is tested against an alternative of bioequivalence, as follows:

$$H_0: GMR \leq 0.80 \text{ or } GMR \geq 1.25$$
$$H_1: 0.80 < GMR < 1.25.$$

In the Reference Scaled Average Bioequivalence (RSABE) approach, the ABE limits are a function, say GMR_{EMA}, of the unknown population within-subject *R* coefficient of variation CV_{WR}, so the hypotheses being tested differ from the standard ones enunciated above:

$$H_0: GMR \leq 1/GMR_{EMA}(CV_{WR}) \text{ or } GMR \geq GMR_{EMA}(CV_{WR})$$
$$H_1: 1/GMR_{EMA}(CV_{WR}) < GMR < GMR_{EMA}(CV_{WR}).$$

If CV_{WR} < 30%, $GMR_{EMA}(CV_{WR}) = 1.25$; so the ABE limits are the usual 0.8–1.25. If CV_{WR} lies between 30% and 50%, the ABE limits grow as $GMR_{EMA}(CV_{WR}) = exp\{k_{EMA}\sqrt{log(CV_{WR}^2+1)}\}$, with $k_{EMA} = 0.76$. Otherwise, from CV_{WR}= 50%, $GMR_{EMA}(CV_{WR}) = 1.4319$; so the ABE limits stay constant at 0.6984 (= 1/1.4319).

A short statement of the EMA testing decision criterion is:

(1) Obtain the *GMR* estimate, $\widehat{GMR} = e^{\hat{\phi}}$, where $\hat{\phi}$ is the estimated formulation effect ϕ, the mean difference of test and reference products of the corresponding log *Cmax* scale;
(2) Point estimate constraint: If $\widehat{GMR}$ is outside the limits 0.8-1.25, do not declare bioequivalence and stop;
(3) Obtain the estimate of the within-subject coefficient of variation of the reference product, $\widehat{CV}_{WR} = \sqrt{e^{\hat{\sigma}_{WR}^2} - 1}$, where $\hat{\sigma}_{WR}^2$ is the estimated value of the reference residual standard deviation in the logarithmic scale;
(4) Obtain the 90% confidence interval for *GMR* around its estimate $\widehat{GMR}$, $CI_{\widehat{GMR}} = e^{[\hat{\emptyset}_L, \hat{\emptyset}_U]}$, where $\hat{\emptyset}_L$ and $\hat{\emptyset}_U$ are the estimated lower and upper limits of the confidence interval in the logarithmic scale, at a confidence level of $1 - 2\alpha$ for α = 0.05
(5) If $CI_{\widehat{GMR}}$ is fully included in the $GMR_{EMA}(\widehat{CV}_{WR})$ limits, declare ABE (reject H_0), otherwise do not declare ABE.

Note that the limits $GMR_{EMA}(\widehat{CV}_{WR})$ are random, not fixed constants like 0.8 or 1.25, since they depend on the random quantity $\widehat{CV}_{WR}$, which is not fixed in advance.

Muñoz et al,[9] among others, showed that the above decision criterion does not adequately control the type I error probability, or false positive rate (say, if bioequivalence is erroneously declared when in fact it does not hold) in the neighborhood of CV_{WR} = 30%.

2.2. Significance level adjustment on the Regulatory EMA scaled approach

As has been previously stated, the 2010 former EMA RSABE procedure does not control completely the type I error probability. To focus on an easy to use method for practitioners, and with chances to be included in the regulations, we considered the method already implemented in the function "scABEL.ad" in the R package PowerTOST.[10] As a consequence of adjusting the significance level, the EMA's scaled method (labeled AdjEMA in the table results) may lose some power. But this (small in general) loss of power is worth because it converts a potentially invalid procedure (with respect to the type I error probability) in a fully correct one.

As a function of the reference coefficient of variation, the type I error probability has only one single maximum at CV_{WR} = 30%. Consequently, though somewhat conservatively, we let the argument "CV" of scABEL.ad at its default value of 0.3. The alternative strategy of estimating the coefficient of variation from data and assigning this (random function of data, unknown in advance) value to the argument CV induces some type I error probability inflation.

In accordance with EMAs Questions & Answers guideline,[11] section 10, the estimation of the required parameters was based on the ANOVA procedure labelled as "Method A" in this document, and not in the intra-subject contrasts, as are for example allowed in the FDA regulation for scaled average bioequivalence.

2.3. Two-stage modified Potvin B and C designs

We consider two adaptive two-stage designs (TSD) with one interim analysis (at the first stage) with N_1 subjects to either (1) establish equivalence early; or (2) stop for futility; or (3) recruit an additional group of N_2 subjects to repeat the bioequivalence assessment at a second stage with $N = N_1 + N_2$ subjects. Each stage is based on a 2×2 crossover balanced *RT/TR* design, and so the within-subject variability CV_W should be estimated by means of the pooled variability of *R* and *T*. Unlike the scaled approach, two-stage hypotheses always rely on the standard fixed limits 0.8–1.25.

Among adaptive approaches to bioequivalence,[15] we focused on those (almost partially) mentioned in regulations, considering two "Pocock-like" variants,[16] as described by Potvin et al. and labelled A, B, C and D.[17] In particular, we studied a Type 1[5] Potvin B method consisting of using the same adjusted α in both stages regardless of whether a study stops in the first stage or proceeds to the second stage (Figure 1), and a Type 2 Potvin C method where an unadjusted α may be used in the first stage, dependent on interim power (Figure 2).

Both methods calculate N_2 as the minimum even number of additional subjects required for having a total sample size of N, which achieves a conditional power of at least 80% for declaring bioequivalence at the second stage. This is conditional on the estimated within-subject coefficient of variation $\widehat{CV}_W$ at the first stage for an assumed true *GMR* of 0.95.

Potvin A was discarded, as it did not adjust the significance boundaries; Potvin D was a more conservative variant of Potvin C, and therefore not recommended because it requires larger average sample sizes than Potvin C.[13]

We propose a modification to the original Potvin B and C algorithms, including two constraints consisting of using a minimum sample size in the second stage (like in other jurisdictions or organizations),[5] and a maximum overall number of 150 subjects enrolled[18,19] in ABE studies, as follows:

- A minimum of $N \geq 1.5N_1$ is required (or $N_2 \geq 0.5N_1$)
- If $N = N_1 + N_2 > 150$, the trial fails and it is stopped at the first stage.

In any case, regardless of the method used, at least 12 evaluable subjects should be included in the first stage.[1,11]

The adjusted significance level of $\alpha = 0.0294$ used by Potvin et al[13,16,17,18] at each stage did not always control the overall type I error rate at a maximum 0.05 (e.g., when using our modified Potvin C algorithm with $N_1 = 12$ and considering a true unknown $CV_W =$ 20%, the false positive rate would be inflated to 0.053). Like in Xu et al,[20] we did look for a significance level by strictly controlling the type I error rate below 0.05, which was useful for our specific modified Potvin B and C methodologies. Because the sponsor is unaware of the true CV_W value, we looked for a significance level which was applicable to a broad set of N_1 and CV_W, $\{N_1/CV_W\}$ (scenarios shown in Section 2.5.).

We used the method implemented in the function "power.2stage" (via non-central *t*-distribution) in the R package Power2Stage. The treatment effect was evaluated at the frontier 1.25, and assuming an expected *GMR* = 0.95 and a target power of 80%.

A short statement for assessing the adjusted significance level, α_{adj}:

(1) Define a grid with a set of $\{N_1/CV_W\}$
(2) Start with an arbitrary, e.g. $\alpha_{adj} = 0.0290$
(3) Obtain the empirical probability of type I error, $Pr\{TIE\}$, over the grid (m = 30,000 simulation trials per scenario). Filter for the scenarios where $Pr\{TIE\}$ is at least 95% of the $max(Pr\{TIE\})$ observed in the grid, let's say $\{N_1/CV_W\}_{TIE \geq P95\%}$
(4) For $\{N_1/CV_W\}_{TIE \geq P95\%}$, find the N_1/CV_W with $max(Pr\{TIE\})$ (m = 1,000,000)
(5) Set up a range of α_j close to the one used before, $\alpha_j \in \{\alpha_{adj} \pm \delta_j\}_{j=1\ldots5}$ (e.g. by δ increments of 0.0001 units). By using the N_1/CV_W associated to $max(Pr\{TIE\})$, estimate the $Pr\{TIE\}$ of all α_j (m = 1,000,000)
(6) Adjust linear $\alpha = g_{lin}(Pr\{TIE\})$ and quadratic $\alpha = g_{quad}(Pr\{TIE\})$ models, with and without the intercept. Choose the model with the lowest Akaike information criterion value (AIC)
(7) Use this model to predict a new α_{adj}, where $\alpha_{adj} = g(0.05)$
(8) Evaluate the entire grid of $\{N_1/CV_W\}$ with this new α_{adj} (m = 1,000,000)

(9) If $Pr\{TIE\} < 0.05$ for all $\{N_1/CV_W\}$, STOP and select this new α_{adj}; Otherwise, start again over with step (4)

As the 2010 EMA guideline uses a Type 1 TSD method,[2] we used the modified Potvin B as the main TSD approach and the modified Potvin C as a sensitive case.

2.5. Simulation methods

The results described in the next sections are based on simulations using 64 bits R and Microsoft R Open. The main outputs are: type I error rate, power and the number of trials stopping at the first stage for the TSD approach. For most scenarios, m = 100,000 datasets were generated, but m = 1,000,000 for those devoted to estimating the most crucial type I error probabilities, i.e., for simulated *GMR*s just on the bioequivalence limit.

In the simulations, we considered all combinations of 3 factors: sample size, true *GMR* and true within-subject variability under the homoscedasticity assumption that $CV_W = CV_{WR} = CV_{WT}$ (from now on, we use CV_W and CV_{WR} interchangeably, provided the assumed simulated homoscedasticity). The sample sizes were N_1 = 12, 18, 24, 30, 36, 48 and 60 subjects for RSABE methods and at the first stage for TSD methods, always considering a balanced design, i.e.: 6, 9, 12, 15, 18, 24 and 30 subjects per sequence. The simulated population *GMR* values were 0.95, 1.00, 1.12, 1.25 and 1.31; with the first three corresponding to scenarios under true bioequivalence (alternative hypothesis), and the last two corresponding to the true non-bioequivalence (null hypothesis). In fact, this statement is exactly true for the TSD approach, where the bioequivalence limits are the constants 0.80–1.25; see the next paragraph for clarification in the RSABE case. Finally, the simulated within-subjects coefficients of variation were 10%, 20%, 25%, 30%, 40%, 50% and 60%. A coefficient of variation of 30% or higher indicates an HVD. Section 3 reports only the results for a subset of the simulated values on sample size, true *GMR*, and true coefficient of variation. In addition, these TSD simulations were done using the "exact" method.

Provided that TSD and RSABE are based on different definitions of bioequivalence, comparing them is quite difficult. In order to have a reference case for comparison, we took the simulated true *GMR* values "on the frontier" of each approach (constant 1.25 in TSD or a function GMR_{EMA} in RSABE for varying simulated CV_{WR} values), which should provide similar proportions of bioequivalence declaration (near 0.05) if both approaches are adequately controlling the user's risk. For *GMR*s that are progressively inside or outside the corresponding bioequivalence regions, these probabilities should also be comparable. To define these concordant simulation scenarios, we reasoned at the logarithmic scale. The constant simulated *GMR* values in the TSD approach are 0.95, 1.00, 1.12, 1.25 and 1.31, and they correspond to formulation effects on the logarithmic scale of –0.0513, 0, 0.1133, 0.2231 and 0.2700, respectively. With respect to the (frontier) 0.2231 value, these formulation effects correspond to proportions λ = –0.230, 0, 0.508, 1 and 1.210, respectively. Then, $\lambda = 1$ refers to values on the frontier, $|\lambda| < 1$ to scenarios of true bioequivalence, and $|\lambda| > 1$ to scenarios of bioinequivalence. Therefore, the same λ value defines concordance in TSD and RSABE scenarios: the population *GMR*s in the original scale were taken as $exp\{\lambda\ 0.2231\}$ in the TSD approaches, and for all simulated CV_{WR} values; while in the RSABE approach, they

were taken as $exp\{\lambda\, 0.2231\}$ for $CV_{WR} < 30\%$, as $exp\left\{\lambda k_{EMA}\sqrt{log(CV_{WR}^2+1)}\right\}$ for CV_{WR} values between 30% and 50%, and as $exp\{\lambda\, 0.3590\}$ for a $CV_{WR} \geq 50\%$.

For simplicity, the simulated *GMR*s in the next sections will always be labeled as 0.95, 1.00, 1.12, 1.25 and 1.31; but it should be remembered that these values in the RSABE case correspond only to the simulated coefficients of variation below 30%.

Following the EMA Questions & Answers guideline,[11] adjusted ANOVA models for analysis of the combined second stage data included the following terms: stage, sequence, interaction sequence*stage, subject nested in sequence*stage, period nested in stage, and formulation.

3. Simulation results

The adjusted significance level predicted for the modified Potvin B was assessed at α_{adj} = 0.0301 at each stage; For the modified Potvin C, the adjusted significance level predicted was assessed at α_{adj} = 0.0280 (Figures 1 and 2).

Both adaptive TSD modified Potvin B and C methods performed similarly in respect to the power achieved and the required median sample size *Me[N]* (Table 1). Because almost all simulated studies required stepping up to a second stage and resulted in large final sample sizes, it was not advisable to start with a too small sample size, like N_1 = 12, in scenarios with high variability ($CV_W \geq 30\%$).

On the other hand, when $N_1 \geq 24$, the global power (including both stages) was at least 80% when variabilities were raised up to 40%. Additionally, those sample sizes increased the likelihood of stopping for bioequivalence at the first stage. For the high value of CV_W = 60%, results were poor, with power always below 80%.

For the RSABE EMA method, a crucial variability value is at the threshold $CV_W = 30\%$, where there is a maximum type I error peak. Table 2 shows that for a true GMR of 1.25 the highest false positive rate is 0.085, confirming the already known risk control problems of the EMA scaled approach. On the other hand, the RSABE adjusted EMA method (AdjEMA) accurately respected the nominal 0.05 level. Both TSD approaches also respected the type I error at 0.05. In addition, for a sample size of N_1 = 24, all methods with a type I error close to the nominal 0.05 level provide satisfactory and similar powers on bioequivalent drugs (*GMR* = 0.95, 1.00, and 1.12). The apparently larger sample sizes required by TSD methods should be relativized: with half periods, they did not double mean size and reached a bioequivalence statement at the first stage in a notable proportion of times (approximately 41%, 47% and 24%).

Figure 3 shows a more comprehensive picture of the extended N_1 and CV_W values for a bioequivalent scenario fixed at *GMR* = 0.95. When N_1 = 12, TSD methods showed higher power than the RSABE adjusted EMA method for $CV_W > 20\%$, requiring relatively larger global sample sizes of *Me[N]* = 44 and around 70 for $CV_W = 30\%$ and 40%, respectively. For N_1 = 24 the RSABE adjusted EMA method showed a similar trend as both TSD methods; and for N_1 = 36, both methods showed power above 80%, for a true CV_W below 60%. For a true $CV_W \geq 60\%$, the power for both TSD

methods seriously suffered from the futility criterion of not allowing studies with more than 150 subjects, though for the RSABE adjusted EMA the power was still above 80%.

Figure 4 explores the power for different true levels of bioequivalence: *GMR* = 0.95, 1.00, and 1.12. It is remarkable that for a true value of *GMR* = 1.12, no methods reached 80% power for any HVD with $CV_W \geq 30\%$.

4. Discussion

Bioequivalence studies are the pivotal clinical studies submitted to regulatory agencies to support the marketing applications of new generic drug products. High levels of within-subject variability make it difficult to assess bioequivalence through standard procedures using reasonable sample sizes, thus delaying treatment. After many years of discussion, some agencies issued regulations describing those methods. In general, their approach is based on bioequivalence limits being scaled as a function of the reference formulation variability. This is the reference scaled average BE (RSABE) approach of the EMA regulation issued in 2010.[2] Although also mentioned in the regulations, adaptive two-stage designs (TSD) are not used nearly as much as the widespread scaling methods, despite having some appealing characteristics. Deciding on the study's experimental design is crucial and must be done in advance (e.g., including it in the study protocol), generally without full knowledge of the within-subject variability. We compared two variants of well-known adaptive methods and an RSABE adjusted (type I error) EMA approach. Both methods showed similar statistical power, but the RSABE adjusted scaled method required less sample size, although at the expense of exposing subjects twice as long as TSD methods. For initial sample sizes of at least 24 subjects, TSDs are a good option to consider, as they have a power of around 80% at the first stage for non-highly variable drugs while at the same time they offer the opportunity for stepping up to the second stage (including additional subjects) for truly bioequivalent products.

Statistical power is used to evaluate the performance of adaptive methodologies in ABE clinical trials. A power of at least 80% is desirable when considering N_1 subjects at the first stage, and assuming an expected but unknown within-subject coefficient of variation, *CVw*. In turn, this is always conditioned to not exceed the overall type I error rate of 0.05 for true bioinequivalent drugs. In our modified Potvin B and C methods, we found adjusted significance levels covering a wide range of N_1 and *CVw* combinations (i.e. α_{adj} = 0.0301 and α_{adj} = 0.0280 at each stage for Potvin B and C, respectively). This is useful to regulators since they can widely rely on the protection of patients against false positive results. However, we understand that for a specific actual (local) N_1 and *CVw* combination, the power might be slightly downgraded, although it is always above 80% in case of true bioequivalence.

Patterson et al[21] explored the sample size that provides 90% power (for true bioequivalent drugs) in case of HVD. They showed that by using 2x2 crossover designs with conventional ABE limits of 0.8-1.25 and *CVw* of 60% or above, the required sample size exceeds 150 subjects (though replicate designs require smaller sample size). Using adaptive designs, we avoid conducting studies with such a large sample size by imposing a futility criterion so that we can stop the trial at an interim look with only N_1 subjects. According to Karalis and Macheras,[19] we added a constraint to the original TSD methods, specifically by not recruiting more than 150 subjects overall. For

example, in the case of a true bioequivalent drug with $0.95 \leq \text{GMR} \leq 1.05$, and for highly variable drugs with an estimated within-subject coefficient of variation above 58% at the interim analysis, the final sample size needed for achieving a power of 80% at the second stage already exceeds 150 subjects. At first glance this constraint represents some global loss of power, but this possibility of cancelling a study for futility may ultimately be considered a positive trait, since the sponsor is unaware of the true treatment effect value during the planning phase, and the overall sample size could unnecessarily soar above this threshold for a scenario of true bioinequivalence. However, from an ethical perspective even starting a study with such a low expected power might be questionable.[22]

Kieser and Rauch[15] and Karalis and Macheras[19] pointed out a potential limitation of the original TSD methods stated by Potvin et al[17] and Montague et al,[13] as although unblinded data are available after the first stage, the knowledge about the estimated *GMR* in the interim analysis is not used for sample size recalculation. We assumed a fixed true treatment effect of *GMR* = 0.95 after the first stage since Cui et al[23] showed that a determination of the second stage sample size based on an interim estimate of the *GMR* can substantially inflate the probability of type I error in most practical situations.

In addition, the expected total sample size *E[N]* is usually used to compare the performance characteristics of different TSD methods. However, by their very nature in TSD, the distribution of total sample sizes N is bimodal, mainly due to the imposition of $N \geq 1.5N_1$. For example, using our modified Potvin B, with $\alpha_{adj} = 0.0301$ at each stage, *GMR* = 0.95, *CVw* = 0.3, N1 = 24, and target power 80%, we obtain a *E[N]* of 40 subjects, but with 24 and 36 subjects having more likelihood of occurrence (Figure 5). As the average is skewed towards two sample values, we believe that the median of *N* is more useful to compare different TSD methods.

In general, regulators allow using adaptive methods, though they usually favor sample size re-estimation procedures that maintain the blinding of the treatment allocations throughout the trial, as shown by Golkowski et al.[24] However, even though both TSD Potvin B and C methods studied in this article assume unblinded data at the interim analysis, the agencies do specifically also recommend using these two TSD methods,[2] as they have demonstrated that they control the type I error rate in a strong way.

So, given that either the RSABE or TSD methods are suitable approaches for ABE studies, we have compared them through the behavior of the type I error rate and its power to facilitate the discussion about which to choose. In terms of power, both approaches perform similarly despite both adaptive methods requiring a higher mean sample size to reach the same power, especially for clearly variable drugs. Nevertheless, they demonstrate suitable power at the first stage in some cases. However, as RSABE relies on replicate designs, double exposure of subjects is needed. The crucial point to consider is the assessment made by sponsors regarding the relative importance of the number of required subjects (an argument favoring the scaled approach) and the exposure of these subjects (which tips the balance in favor of the TSD approach).

The applicability of the TSD approaches is essentially the same as the classical approach, in that they have the same *RT/TR* design and fixed standard limits.[25] The RSABE approaches (with type I error adjustment) are appropriate for drugs with low to moderate variability, because dose-to-dose variability within a patient is comparable to

the width of the criteria. However, with HVD, dose-to-dose variability within a patient is greater than the width of the standard criteria, and it is usually characterized by flat dose response curves and wide safety margins. Therefore, broadening the acceptance limits in the RSABE approach is at the very least controversial, since clinically sound criteria should be used to clearly prove if a greater difference in *Cmax* (and also in *AUC* for the FDA) is irrelevant.

In conclusion, the RSABE approach is well powered and usually requires enrolling fewer patients than adaptive TSD methods, even though scaling the ABE limits ultimately depends on additional clinical judgment. For HVD in general, samples of 36 subjects provided well-powered studies using RSABE methods. As there is a considerable chance of declaring ABE at the first stage in adaptive approaches, sponsors should consider them because they imply less subject exposure and less treatment duration.

Acknowledgments

This research is partially supported by the grant MTM2015-64465-C2-1-R (MINECO/FEDER) from the Ministerio de Economia y Competitividad (Spain) and by the grant 2014 SGR 464 from the Generalitat de Catalunya.

We would like to thank the reviewers who identified areas of the manuscript that needed corrections or modifications.

References

1. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research. Guidance for Industry: Bioavailability and Bioequivalence Studies submitted in NDAs or INDs - General considerations. http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm389370.pdf. Published March 2014. Accessed October 18, 2016.

2. European Medicines Agency. Guideline on the investigation of bioequivalence. CPMP/EWP/QWP/1401/ 98 Rev.1/Corr. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/01/WC500070039.pdf. Published January 2010. Accessed October 18, 2016.

3. Tothfalusi L, Endrenyi L, Garcia Arieta A. Evaluation of bioequivalence for highly variable drugs with scaled average bioequivalence. Clin Pharmacokinet. 2009;48(11):725-743.

4. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research. Guidance for Industry: Statistical Approaches to Establishing Bioequivalence. https://www.fda.gov/downloads/drugs/guidances/ucm070244.pdf. Published January 2001. Accessed October 18, 2016.

5. Schütz H. Two-stage designs in bioequivalence trials. Eur J Clin Pharmacol. 2015;71(3):271-281.

6. Davit BM, Conner DP, Fabian-Fritsch B, et al. Highly variable drugs: observations from bioequivalence data submitted to the FDA for new generic drug applications. AAPS J. 2008;10(1):148-156.

7. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research. Guidance for Industry: Bioequivalence Studies with Pharmacokinetic Endpoints for Drugs Submitted Under an ANDA. http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm377465.pdf. Published December 2013. Accessed October 18, 2016.

8. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research. Draft Guidance on Progesterone. http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm209294.pdf. Published February 2011. Accessed October 18, 2016.

9. Muñoz J, Alcaide D, Ocaña J. Consumer's risk in the EMA and FDA regulatory approaches for bioequivalence in highly variable drugs. Stat Med. 2016;35(12):1933-1943.

10. Labes D, Schütz H. Inflation of Type I error in the evaluation of scaled average bioequivalence, and a method for its control. Pharm Res. 2016;33(11):1-10.

11. European Medicines Agency. Questions & Answers: Positions on specific questions addressed to the pharmacokineticsworking party.

http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002963.pdf. Published November 2015. Accessed October 18, 2016.

12. Bandyopadhyay N, Dragalin V. Implementation of an adaptive group sequential design in a bioequivalence study. Pharm Stat. 2007;6(2):115-122.

13. Montague TH, Potvin D, DiLiberti CE, Hauck WW, Parr AF, Schuirmann DJ. Additional results for "Sequential design approaches for bioequivalence studies with crossover designs". Pharm Stat. 2012;11(1):8-13.

14. Davit B, Braddy AC, Conner DP, Yu LX. International guidelines for bioequivalence of systemically available orally administered generic drug products: a survey of similarities and differences. AAPS J. 2013;15(4):974-990.

15. Kieser M, Rauch G. Two-stage designs for cross-over bioequivalence trials. Stat Med. 2015;34(16):2403-2416.

16. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika. 1977;64(2):191-199.

17. Potvin D, DiLiberti CE, Hauck WW, Parr AF, Schuirmann DJ, Smith RA. Sequential design approaches for bioequivalence studies with crossover designs. Pharm Stat. 2008;7(4):245-262.

18. Karalis V, Macheras P. On the statistical model of the two-stage designs in bioequivalence assessment. J Pharm Pharmacol. 2014;66(1):48-52.

19. Karalis V, Macheras P. An insight into the properties of a two-stage design in bioequivalence studies. Pharm Res. 2013;30(7):1824-1835.

20. Xu J, Audet C, DiLiberti CE, et al. Optimal adaptive sequential designs for crossover bioequivalence studies. Pharm Stat. 2016;15(1):15-27.

21. Patterson SD, Zariffa N, Montague TH, Howland K. Non-traditional study designs to demonstrate average bioequivalence for highly variable drug products. Eur J Clin Pharmacol. 2001;57(9):663-70.

22. Fuglsang A. Futility rules in bioequivalence trials with sequential designs. AAPS J. 2014;16(1):79-82.

23. Cui L, Hung HMJ, Wang S-J. Modification of sample size in group sequential clinical trials. Biometrics. 1999;55(3):853-857.

24. Golkowski D, Friede T, Kieser M. Blinded sample size re-estimation in crossover bioequivalence trials. Pharm Stat. 2014;13(3):157-162.

25. European Generic Medicines Association. Revised EMA Bioequivalence Guideline: Questions and Answers. Summary of the discussions held at the 3rd symposium on bioequivalence. http://www.medicinesforeurope.com/wp-content/uploads/2016/03/EGA_BEQ_QA_WEB_QA_1_32.pdf. Published June

2010. Accessed October 18, 2016.

Table 1. Two-stage design (TSD) modified Potvin B and C: Bioequivalence, sample size, and percentage of studies stepping up to second stage for true *GMR* = 0.95, and under different fixed N_1 and a true *CVw*

		Modified Potvin B								Modified Potvin C							
Fixed a priori		ABE		Step to St2	*N*					ABE		Step to St2	*N*				
N_1	True *CVw*	% St1	% St1+St2	%	Min	5%	Me	95%	Max	% St1	% St1+St2	%	Min	5%	Me	95%	Max
12	20	41.92	85.00	55.69	12	12	18	40	104	41.56	84.76	54.44	12	12	18	40	106
12	30	7.03	78.61	92.71	12	12	44	84	150	6.40	78.34	93.05	12	12	44	84	150
12	40	1.03	71.65	95.68	12	22	70	128	150	0.90	70.96	95.28	12	20	72	130	150
12	60	0.05	29.43	51.00	12	12	44	142	150	0.05	27.76	49.06	12	12	12	142	150
24	20	83.76	90.16	8.20	24	24	24	36	62	87.89	91.19	4.22	24	24	24	24	64
24	30	41.86	83.86	57.47	24	24	36	70	138	40.47	83.38	57.69	24	24	38	72	140
24	40	10.12	79.79	89.45	24	24	76	118	150	8.93	79.44	90.49	24	24	78	120	150
24	60	0.19	31.19	46.47	24	24	24	146	150	0.15	28.83	43.59	24	24	24	146	150
36	20	95.68	95.75	0.07	36	36	36	36	54	97.51	97.51	0.01	36	36	36	36	54
36	30	68.13	87.23	28.33	36	36	36	60	120	69.94	85.77	22.95	36	36	36	62	124
36	40	34.32	82.42	65.54	36	36	68	110	150	32.40	82.14	67.16	36	36	72	112	150
36	60	1.53	31.28	42.66	36	36	36	146	150	1.20	28.35	39.37	36	36	36	146	150

ABE, average bioequivalence; TSD, two-stage design; *GMR*, geometric mean ratio; N_1, initial and fixed sample size (Stage 1); *CVw*, within-subject coefficient of variation; %St1, proportion of simulations declaring bioequivalence at Stage 1; %St1+St2, cumulative proportion of simulations declaring ABE at Stage 2, Step up to St2, proportion of simulations requiring stepping up from Stage1 to Stage 2; Min, min of *N*; 5%, percentile 5 of *N*; Me, median of N; 95%, percentile 95 of *N;* Max, max of *N*;

Table 2. Probability of bioequivalence acceptance according to the regulatory reference scaled ABE (RSABE) EMA and an adjusted EMA method compared to two-stage designs (TSD) modified Potvin B and C (true *CVw* = 30%)

		Probability ABE acceptance			Type I error	
		True *GMR*				
	Method	**0.95**	**1.00**	**1.12**	**1.25**	**1.31**
RSABE method	Regulatory EMA (N_1 = 24)	0.896	0.963	0.631	0.085	0.021
	AdjEMA (N_1 = 24)	0.864	0.948	0.559	0.050	0.009
TSD method	Modified Potvin B (N_1 = 24 at Stage 1)	0.419	0.484	0.242	0.029	0.008
	Modified Potvin B (Stage 1 + Stage 2 with 36≤*N*≤150)	0.839	0.926	0.527	0.050	0.012
	Modified Potvin C (N_1 = 24 at Stage 1)	0.405	0.468	0.236	0.030	0.009
	Modified Potvin C (Stage 1 + Stage 2 with 36≤*N*≤150)	0.834	0.922	0.519	0.048	0.012

ABE, average bioequivalence; RSABE, reference scaled average bioequivalence; TSD, two-stage design; *GMR*, geometric mean ratio; *CVw*, within-subject coefficient of variation; N_1, initial and fixed sample size fixed at 24 subjects (Stage 1 with modified Potvin B and C); Regulatory EMA, regulatory European Medicines Agency approach; AdjEMA, adjusted EMA type I error

Legends

Figure 1. Type 1 TSD Modified Potvin B algorithm

Adapted from the figure depicted in detail by Montague et al,[13] with the restriction of Karalis and Macheras[18] of not including more than 150 subjects and $N \geq 1.5N_1$;
ABE, average bioequivalence; N_1, Initial fixed sample size; N_2, the additional number of subjects recruited at Stage 2; *GMR*, assumed geometric mean ratio; $\widehat{CV}_W$, estimated within-subject coefficient of variation

Figure 2. Type 2 TSD Modified Potvin C algorithm

Adapted from the figure depicted in detail by Montague et al,[13] with the restriction of Karalis and Macheras[18] of not including more than 150 subjects and $N \geq 1.5N_1$;
ABE, average bioequivalence; N_1, Initial fixed sample size; N_2, the additional number of subjects recruited at Stage 2; *GMR*, assumed geometric mean ratio; $\widehat{CV}_W$, estimated within-subject coefficient of variation

Figure 3. Bioequivalence acceptance of the adjusted reference scaled ABE (RSABE) EMA method and two-stage designs (TSD) modified Potvin B and C at stages 1 and 2, for a true *GMR* of 0.95, and a progressive increase of the within-subject variability

ABE, average bioequivalence; RSABE, reference scaled average bioequivalence; TSD, two-stage design; *GMR*, geometric mean ratio; HVD, highly variable drugs; N_1, initial and fixed sample size used for the modified EMA method and both TSD methods at Stage1; *CVw*, within-subject coefficient of variation; *Me[N]*, TSD media total sample size at Stage 2 (beside the squares in the figure); AdjEMA, type I error adjusted EMA method

Figure 4. Bioequivalence acceptance of the adjusted reference scaled ABE (RSABE) EMA method and two-stage designs (TSD) modified Potvin B for different levels of true bioequivalence and a progressive increase in the within-subject variability

ABE, average bioequivalence, RSABE, reference scaled average bioequivalence; TSD, two-stage design; HVD, highly variable drugs; N_1, initial and fixed sample size (EMA method); *GMR*, geometric mean ratio; *CVw*, within-subject coefficient of variation; *Me[N]*, TSD median total sample size (beside the squares in the figure); AdjEMA, type I error adjusted EMA

Figure 5. Type 1 TSD modified Potvin B distribution of *N* (Stag1 + Stage 2) *GMR*=0.95*; CVw*=30%; N_1=24; α_{adj} =0.03018396; *P*=0.8; *m*=1,000,000 simulations

GMR, true geometric mean ratio; *CVw*, true within-subject coefficient of variation; N_1, Initial fixed sample size; N_2, the additional number of subjects enrolled at stage 2; $N=N_1+N_2$, total sample size (stage 1 + stage 2); α_{adj}, significance level used in each stage; *P*. target power; *m*, number of simulations

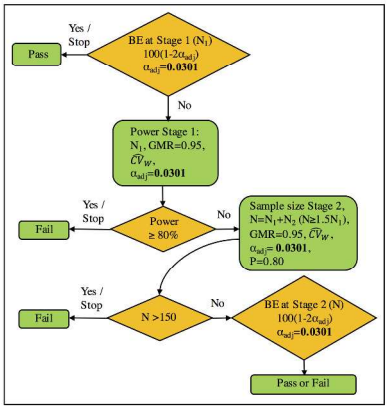

Figure 1. Type 1 TSD Modified Potvin B algorithm

312x220mm (300 x 300 DPI)

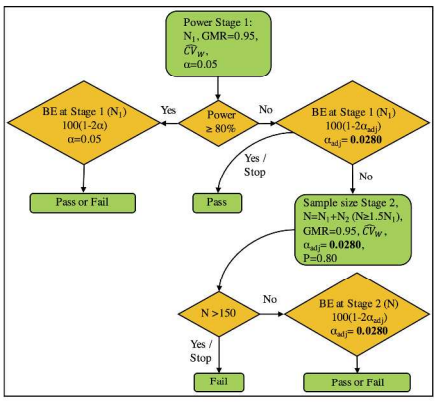

Figure 2. Type 2 TSD Modified Potvin C algorithm

312x220mm (300 x 300 DPI)

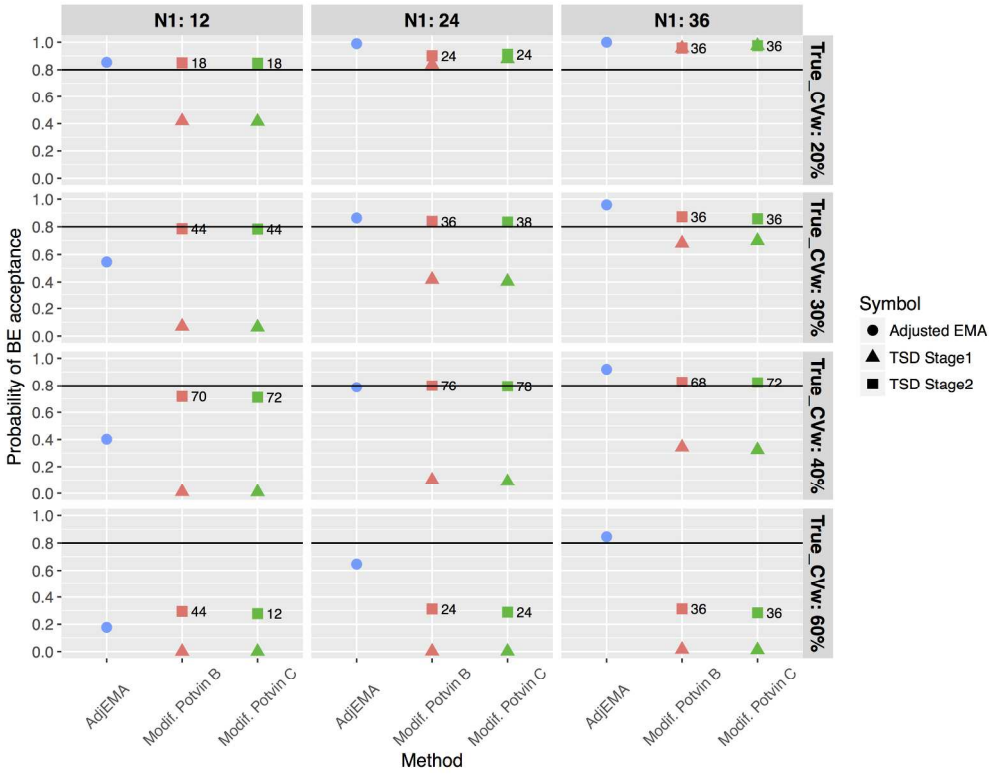

Figure 3. Bioequivalence acceptance of the adjusted reference scaled ABE (RSABE) EMA method and two-stage designs (TSD) modified Potvin B and C at stages 1 and 2, for a true GMR of 0.95, and a progressive increase of the within-subject variability

228x177mm (300 x 300 DPI)

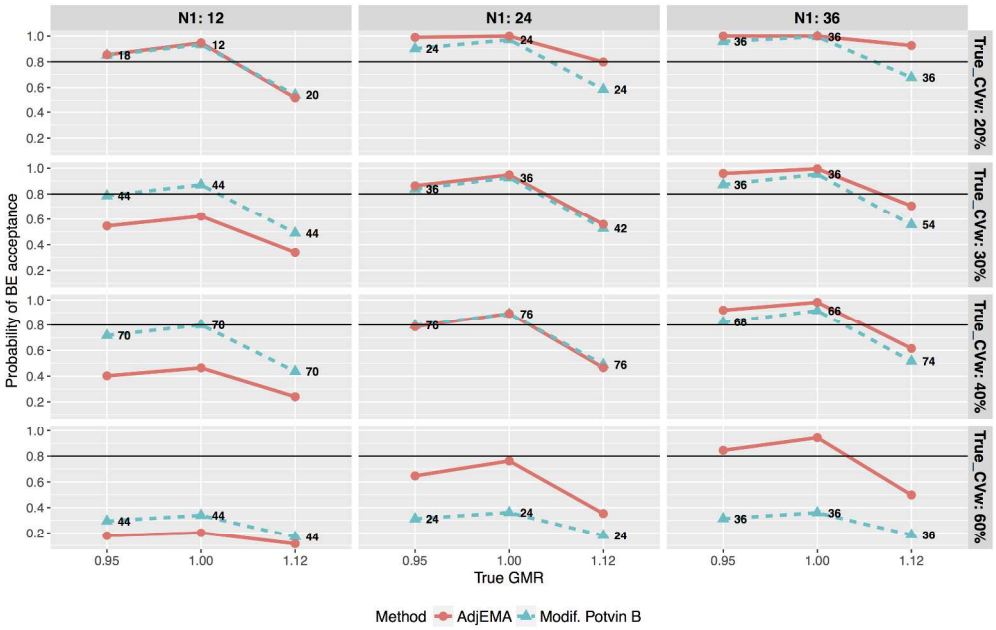

Figure 4. Bioequivalence acceptance of the adjusted reference scaled ABE (RSABE) EMA method and two-stage designs (TSD) modified Potvin B for different levels of true bioequivalence and a progressive increase in the within-subject variability

276x177mm (300 x 300 DPI)

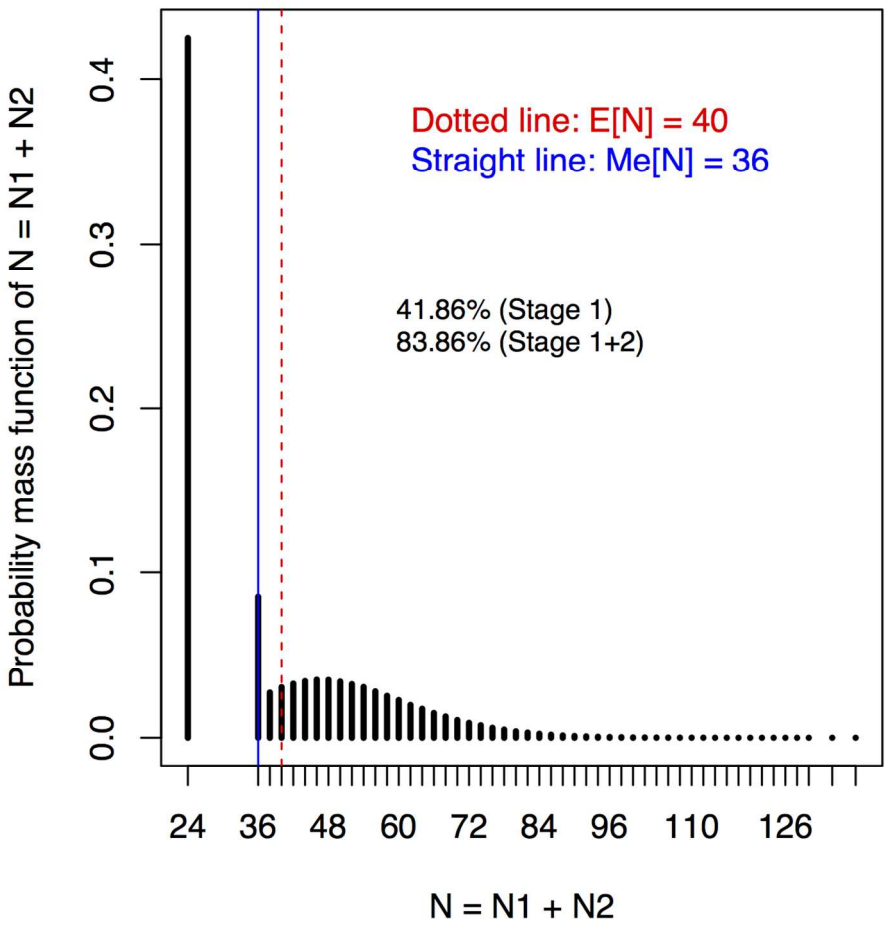

Figure 5. Type 1 TSD modified Potvin B distribution of N (Stag1 + Stage 2)
GMR=0.95; CVw=30%; N1=24; αadj =0.03018396; P=0.8; m=1,000,000 simulations

127x126mm (300 x 300 DPI)